

caBIG

Strategic Planning Face to Face Meeting VCDE Progress Report Philadelphia, PA – Fox Chase Cancer Center November 8-9, 2004

Outline of Presentation:

- ▶ 1. Purpose of Workspace
- 2. Year 1 Goals
- > 3. Accomplishments
- ▶ 4. SWOT Analysis
- ▶ 5. Discussion



1. Purpose of VCDE Workspace

Role:

- responsible for the evaluation and development of software and tools to support the integration and implementation of common data elements (CDEs) and controlled vocabulary content
- responsible for the identification and evaluation of data standards and data quality indicators
- coordinate with Architecture to address compatibility of infrastructure and promote fundamental vocabulary and data standards

Developers:

- Hawaii, Jackson Labs, Mayo
- Adopters:
 - None originally or currently
- Working Group Members:
 - Albert Einstein, Fred Hutch, Pittsburgh, UC Davis





2. Year 1 Goals (Current Projects)

- Plan for governance of new vocabulary and common data element content
- Establish V-CDE Mentoring Teams for domain workspace pilot projects
- Recommendation of external data standards
- Convene regular Face to Face Meetings of the V-CDE Workspace and joint meetings with the Architecture Workspace to address caBIG cross cutting issues
 - First F2F October 27 & 28th in Chicago
 - To be detailed in section 4 of this summary
- Implement Vocabulary/CDE Workspace developer projects
 - To be discussed next (section 3)





Assignment of Development Task Orders (2 of 3)

Hawaii:

- Nutritional ontology for epidemiology studies
 - Work with NCICB EVS group
 - Publish internal dietary databases as caBIG vocabulary CDEs in caDSR
 - Harmonize w/ other databases (USDA, AgOrgUN)

Jackson Labs:

- Mouse Ontology
 - Mapping and harmonization of anatomic descriptors used for murine and human models
 - caBIGify Mouse Phenome Database, the Mouse Tumor Biology database and the Mouse Genome Informatics consortium database (MGI)





Assignment of Development Task Orders (3 of 3)

Mayo:

- LexGRID distributed network of terminology resources
 - Bridge terminologies and ontologies with a common set of tools, formats and update mechanisms
 - Standards access methods and formats published and openly available
 - Tools standards based tools readily available
 - Content commonly used terminologies available for access and download
 - Accessible through a set of common application protocol interfaces (API), it is web accessible and locally extendable





Development of "Mentoring" Teams

▶ Role of Mentoring Teams:

- provide support and guidance to the various Domain Workspaces in the development, administration and utilization of CDEs
 - whether UML model-driven or case report form-driven, integrate via controlled vocabulary
 - priority, is the support required by the developer project teams
- Architecture Fred Hutch (caGRID), Mayo (caGRID),
- Clinical Trials Mayo (Protocol/Billing/Interfaces/ Compatibility/Metadata), UC-Davis (AER/Billing/CDUS/ Metadata)
- ICR Albert Einstein (Pathways), Hawaii (Data Analy & Stats/caArray/Proteomics), Fred Hutch (Translation Tools/Gene Annotation),
- Tissue Banks & Pathology Tools U Pitt





Progress Report & Action Items – Data Standards Arch/VCDE F2F Meeting Chicago 10/27-28

Discussion Items:

- general agreement to caBIG data standards governance process with minor modifications based on feedback
- V-CDE will consider procedures for notification of caBIG members regarding changes to data standards

Action Items:

- Kathleen Gundry will provide V-CDE workspace with advanced copies of data standards under consideration by the NCI Context Administrators as possible template for data standards submission package
- V-CDE to develop detailed process for review, approval, and maintenance of data standards.





Progress Report & Action Items – HL7 Arch/VCDE F2F Meeting Chicago 10/27-28

Discussion Items:

- HL7 does not satisfy all data messaging requirements for caBIG, but may apply to clinical data messaging
- HL7 data types may be applicable outside of the clinical domain
- HL7 V3 is not in wide use and V2 will need to be supported, but Workspace developers should be encouraged to consider V3 because of semantic richness
- Final decisions on applicability of HL7 requires further study

Action Items:

 HL7 v3 messaging use cases need to be defined by Clinical Trials, TB& PT and ICR workspaces





Progress Report & Action Items – Metadata Arch/VCDE F2F Meeting Chicago 10/27-28

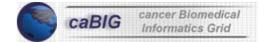
Discussion Items:

- Metadata should be a separate service provided by any node on the Grid.
- Metadata will be represented in a multi-tier fashion.
- The first tier will be an XML schema describing the structure of the data object.
- The next tier will contain the data object's metadata from the caDSR. The proposed representation is an XML schema. We propose conducting a review in the next 2-3 weeks to determine if XML schema is sufficient.
- The final tier will contain the semantic information represented in OWL/OWL-S.

Action Items:

 A minimal set of metadata needs to be identified for data and analytical service providers. The caGRID phase I white paper suggests a starting point that can be used to devise these lists





3. Accomplishments (optional depends on Architecture report) Progress Report & Action Items – Provenance Arch/VCDE F2F Meeting Chicago 10/27-28

Discussion Items:

- caBIG should adopt a hierarchical provenance model
- provenance information should be attached to any persistent data store
- the data store is only required to provide provenance information to its immediate source of information, optional to go back further or original source
- the provider determines the type of information that constitutes its provenance - should be sufficient to find the source information and understand the transformation that was performed on the data
- Analytical services should provide provenance information as the data passes through workflows

Action Items:

 A single group should be formed to work on ID, versioning and provenance.





4. SWOT Analysis-Strengths + Weaknesses-NOTE: MJB Opinion

Strengths:

- Strongest part of NCICB "portfolio" of applications and toolkits appears to be EVS
- Many skilled parties in NCICB, contractors in this area
- Strong commitment (resources) evident
- Strong team in VCDE space
- Mentoring teams is a very valuable addition (need to execute)

Weaknesses:

- caDSR is still not deployed fully (yet)
- Training has been complicated and not fully thought through
- Old CDEs in caDSR are "forms based" and little object model level software deployed in caDSR

cancer Biomedical

- No adopters apparent yet (key weakness, need to balance)
- Not enough emphasis on clinical and translation needs apparent at F2F meeting of Arch/VCDE
- Not enough decisions (but much healthy discussion)



4. SWOT Analysis - Opportunities & Threats-NOTE: MJB Opinion

Opportunities:

- Strengthen resources supporting caDSR
- Publicize early 'heroes' who do end to end integration
- Couple developers with adopters with domain experience NOW
- Champion mentoring teams but to be productive with caDSR and CDE curation ASAP
- Funding growth of caBIG is phenomenal and on track (50% increase for next year)

Threats:

- Cancer Centers program still skeptical with need to see results
- Training program has left the gate slowly (new team in place)
- Architecture moving too slowly could significantly hinder caBIG effort
- Culture of "for profit" software development vs academic mindset
- ICR space will consume significant VCDE resources
 - Is VCDE appropriately resourced?





5. Discussion

Outline of Presentation:

- ▶ 1. Purpose of Workspace
- 2. Year 1 Goals
- > 3. Accomplishments
- 4. SWOT Analysis



